Statistics

My dad works at NASA. I asked him if he wanted to get a PhD with me so we could be study buddies. He asked what I was studying. "Math," I said, "and statistics is first!" He said, "Eww! No thanks! I have always called that Sadistics!" So, in honor of my dad we will from now on call this:

Sadistics

Some Definitions

Population: The entire group that you're collecting information about.

Sample: A subset of the population taken. We do this so it's a smaller, more manageable piece of data or especially when the entire population cannot be surveyed.

Representativeness: Characteristics of a sample have to reflect the entire population, not just part of it.

Mean: Also called the average. It's the sum of all values in the sample divided by the number of values in the sample or population. Less importantly, μ is for populations and x̄ is for samples.

Median: If you arrange the values from lowest to highest, the median is the middle one (or middle two if there is an even number).

Variance: Measures how far away a value is from the mean.

For populations:

$$\sigma^2 = \frac{\sum(x-\mu)^2}{n}$$

For samples:

$$s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

Why on earth are we doing n-1 for samples? There's a math proof to explain this. For now, just remember to do it.

Standard Deviation: Square root of the variance

Covariance: The joint variability of two random variables

For populations:

$$Cov_{x,y} = \frac{\sum(x_i-\mu_x)(y_i-\mu_y)}{n}$$

For samples:

$$Cov_{x,y} = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{n-1}$$

Correlation Coefficient: Measures the strength of a linear relationship between two variables. Your answer should be between -1 and 1.

$$CorrCoeff = \frac{Covariance}{\sqrt{(Variance\ of\ x)(Variance\ of\ y)}}$$

Standard Error: The extent that a statistic changes from sample to sample.

For samples with equal variances:

$$\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

For samples with unequal variances:

$$\sqrt{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)}$$

CONFUSION MATRIX AKA Error Matrix: A table that summarizes the performance of a classification measurement.

	PREDICTED NO	**PREDICTED YES**	Total
ACTUAL NO	True Negative	False Positive	
ACTUAL YES	False Negative	True Positive	
Total			

True Positive = A fire alarm goes off and there's a fire

False Positive = A fire alarm goes off and there's no fire

False Negative = No alarm, but there's a fire

True Negative = No alarm and there's no fire

Introductory Probability Theory

I went a long way in statistics before someone told me that probability is a function. So I say to you… YO!! It's a function!

Observed Probability: Estimated probability based on observation, not experimentation.

$$P_{Observed}(A) = \frac{Number\ of\ times\ "A"\ Occured}{Number\ of\ times\ test\ was\ repeated}$$

Classical Probability: Based on the chance of something occurring. Each element or "event" must have an equal chance of occurring

$$P_C(A) = \frac{Number\ of\ ways\ "A"\ could\ occur}{Number\ of\ simple\ events\ AKA\ outcomes}$$

ADDITION RULE OF PROBABILITY: Finding the probability of two or more events occurring in the same trial (at the same time). It's also known as a **Compound Event,** or event that joins two or more simple events.

DISJOINT: Mutually exclusive / cannot happen at the same time. If the events are mutually exclusive:

$$P(A \cup B) = P(A) + P(B) - \cancel{P(A \cap B)}^{\,0}$$

JOINT: They can happen at the same time. If the events are NOT mutually exclusive:

$$P(A\ \cup\ B) = P(A) + P(B) - P(A \cap B)$$

When you do a problem about this, draw out the confusion matrix so that it's easier to see if the event is mutually exclusive or not.

Example: A court has the following statistics.

	DIDN'T DO IT	**DID THE CRIME**	Total
FOUND GUILTY	(False Positive) 11	(True Positive) 72	83
FOUND NOT GUILTY	(True Negative) 85	(False Negative) 9	94
Total	96	81	177

What is the probability of randomly selecting a person who is either guilty or did the crime? $\frac{Guilty\ or\ did\ the\ crime}{Total} = \frac{72+11+9=92}{177} \approx 52\%$

Example: What is the probability of being blonde or female?

	FEMALE	NOT FEMALE	Total
BLONDE	12	16	28
NOT BLONDE	20	25	45
Total	32	41	73

What is the probability of randomly selecting a person who is either blonde or a female?

$$\frac{Blonde\ female\ or\ bot}{Total} = \frac{12+16+20=48}{73} \approx 65.8\%$$

BAYES THEOREM OF CONDITIONAL PROBABILITY: The probability of event A occurring given that event B has already happened. Or in other words, the probability of A happening on the condition that event B happened.

Now there's another consideration here. Replacement. If we're drawing out cards, the probability will change depending on whether or not the first card was removed or returned (replaced) back into the deck.

Drawing one card and then putting it aside… there's only 51 cards left instead of 52.

Dependent Events, or those that do not have Replacement: The occurrence of one event DOES affect the occurrence of a subsequent event. All non-independent events are dependent. So, here's the formula:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B|A) \cdot P(A)}{P(B)}$$

Independent Events, or those that have Replacement: The occurrence of one event does not affect the occurrence of a subsequent event. So, with the formula above, some of it becomes irrelevant and therefore cancels:

$$P(B|A) = \frac{\cancel{P(A|B)}^{0} \cdot \cancel{P(A)}^{0}}{\cancel{P(B)}_{0}} = P(B)$$

MULTIPLICATION RULE OF PROBABILTY: The probability of two events happening in successive trials. Event "A" happens, and THEN event "B" happens in the next trial.

$$P(A \cap B) = P(A) \cdot P(B|A)$$

You can't go wrong just remembering this formula, but recall that for independent events, the conditionality will cancel out and you'll get just $P(A \cap B) = P(A) \cdot P(B)$

Example: You're stalking me and you need to know the probability of answering these two questions correctly so you can properly follow me home:

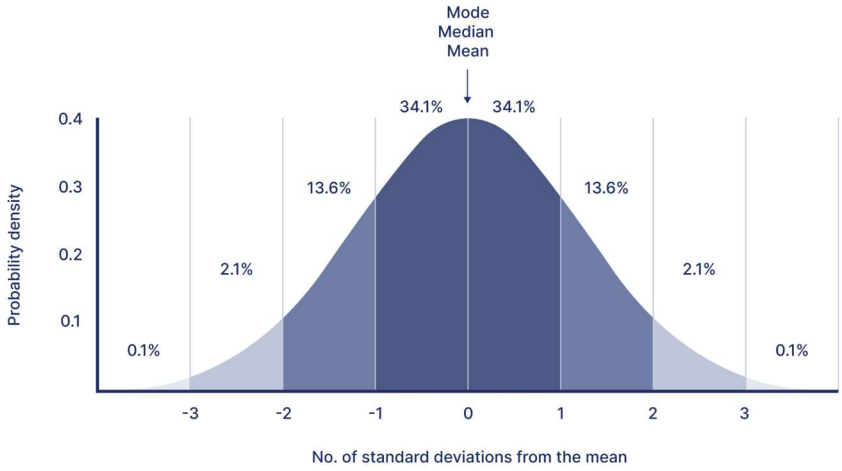
1) True or false: I drive a luxury car
2) Where do I live?
    a. The middle-class suburb with track homes
    b. University area
    c. The gang-banger neighborhood
    d. The fancy mansions and horses neighborhood

What is the probability of getting both of these questions correct?

$$\frac{P(Selecting\ 1\ correctly)}{2\ choices} \times \frac{P(Selecting\ 2\ correctly)}{4\ choices} = \frac{1}{2} \times \frac{1}{4} = \frac{1}{8}$$

Distribution

Standard Normal Distribution: Also called Gaussian distribution. Where values fit a bell curve. 68% of values will fall within 1 standard deviation. 95% fall within two standard deviations. 99.7% fall within three standard deviations.



Outlier: Anything outside 2 standard deviations.

Z-SCORE: How many standard deviations away from the mean something is. Something within 1 standard deviation has a z-score less than 1. Helps us compare the variation in two different samples or populations.

For populations:

$$z = \frac{x - \mu}{\sigma}$$

For samples:

$$z = \frac{x - \bar{x}}{s}$$

Hot Tip: Converting from Z-Score to Area

When looking up Z-Score on a table or using Inverse Norm on a calculator, it gives you the area to the **LEFT** of that point on the curve.

If you need the area to the right, you'll use a different X value.

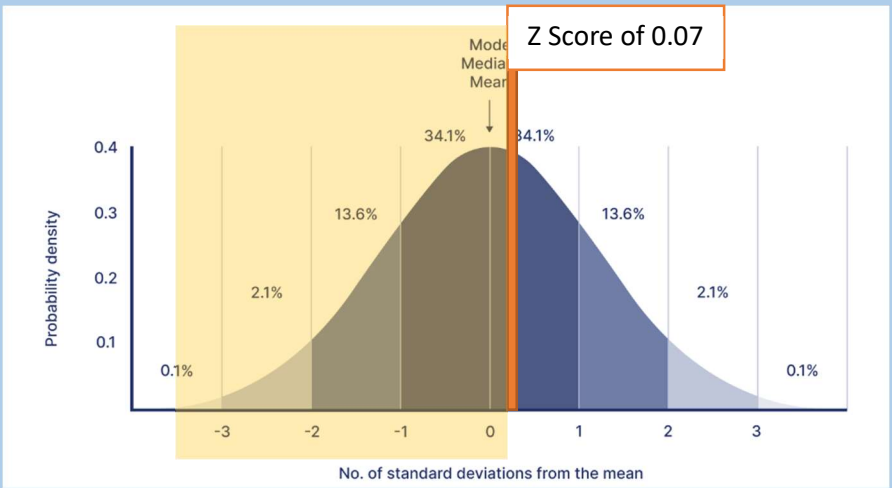
Example: A population of women has a mean weight of 172 pounds and a standard deviation of 29 pounds. What is the probability that a randomly selected woman from this population will have a weight less than 174 pounds?

First, we need to translate this whole shit into something meaningful. First, we need the Z-score.

$$Let\ x = 174,\ \ \mu = 172,\ \ \sigma = 29$$

$$z = \frac{174 - 172}{29} = \frac{2}{29} \approx 0.07$$

Okay so what the f does that mean?



Look this up on a Z-Score table, and you'll see that this matches with 0.5279, or a 52.8% chance, which you can see shaded as the area of the curve overlapping the yellow region.

Z-scores can also be used to find data values.

Example: IQ is normally distributed with a mean of 100 and a standard deviation of 15. What percentage of people have an IQ between 85 and 125?

$$Let\ \mu = 100,\ \ \sigma = 15,\ \ x_1 = 85,\ \ \ x_2 = 125$$

$$z_1 = \frac{85 - 100}{15} = -1.00\,, \quad z_2 = \frac{125 - 100}{15} = 1.67$$

Look them up on a z-score chart. Note: even though you have a negative z-score, you cannot have a negative area or negative probability.

You should get an answer of: 0.7938 so, there is a 79.38% chance.

Skewness: A measure of how abnormally shaped a normal distribution curve is.

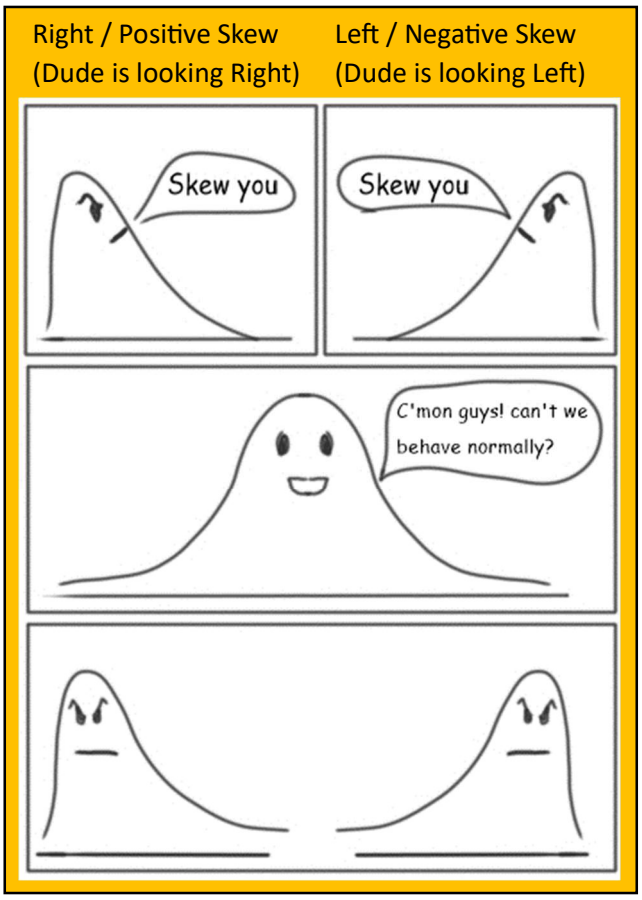
$$Skewness = \frac{\sum_i^N (X_i - \bar{x})^3}{\sigma^3 (N-1)}$$

Kurtosis: A measure of the "tailedness" of the probability distribution of a random value.

$$Kurtosis = \frac{\mu^4}{\sigma^4}$$

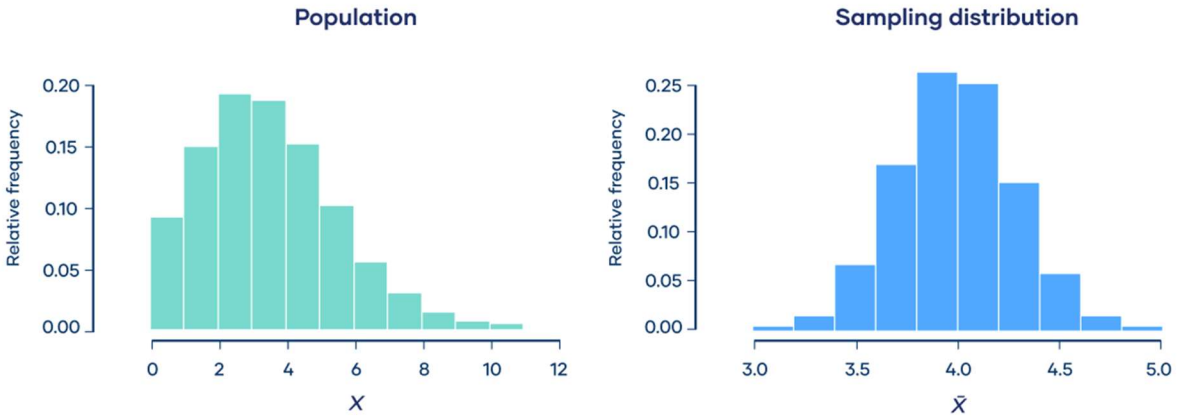
Interquartile Range: With the median, we divided the data in half… this time you're dividing into 4 groups.

IQR 1 = Half of the median;
IQR 2 = Median;
IQR 3 = Median plus IQR 1



CENTRAL LIMIT THEOREM: Regardless of the distribution of the population the sampling distribution of the mean will always be normally distributed, as long as the sample size is large enough.

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad Where: n = sample\ size$$



Cool. What do we use this for? We can use a sample to estimate a population size.

How many samples of size "n" are possible out of a population with size "N"?

It's a combination. NCn. But that number can be VERY big, and you don't want to take all those samples. But some smartie pants out there figured this out:

Sampling Distribution: The average of sample means can be considered equivalent to the population mean.

If $n > 30$ the sampling distribution is normally distributed, and your sample is big enough. Otherwise, you might need a bigger or more representative sample.

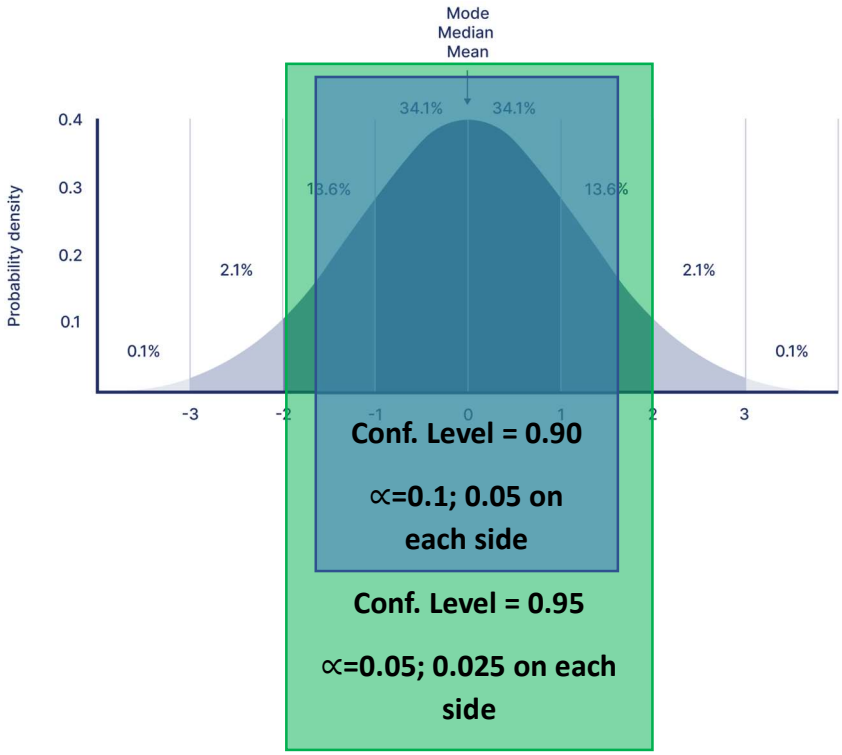
CONFIDENCE INTERVAL: An estimate of population proportion from a sample proportion. It's a range of numbers used to estimate the population parameter.

Here's what you need:

1) Random sample
2) Confidence level – how confident you are that the actual value of the population parameter will be inside the interval.

$1 - \alpha$: Alpha is the complement of a confidence level.

Most common confidence levels: 90%, 95%, or 99%. So, for these levels, the alphas are: 10%, 5%, and 1% respectively.



POINT ESTIMATE: One value from a sample that you use to approximate a population's parameter.

$$\hat{p}\ is\ used\ as\ a\ point\ estimate\ for\ p.$$

$Let: p$ represent the population proportion of success
$\hat{p}$ represent the sample proportion of success
$\hat{q}$ represent the sample proportion of failure

$$\hat{p} = \frac{successes}{trials} = \frac{x}{n}, \qquad \hat{q} = 1 - \hat{p}$$

CRITICAL VALUE: a z-score that separates the "likely" region from the "unlikely" region. Use your table to find z-score from the area.

Area (Confidence Level)	Z-Scores (Critical Values)	Alpha / 2
0.90	-1.645, 1.645	0.05
0.95	-1.96, 1.96	0.025
0.99	-2.575, 2.575	0.005